# Final Report

Tong Zou

March 21, 2018

**Abstract**

In this study the goal is (i). to draw inference on association between odds of completing a 4-year dementia prevention study and type of study partner who the subject participate along with, (ii). to build a predictive model evaluating probability of subject completing the study. The data used is a 4-year dementia prevention study with 644 participants from 39 sites scattered across the United States. Conclusion is made that similar in age, gender, education of both participant and partner, history of smoking , medical history of cardiovascular disease and cancer, given no sign of dementia, the estimated relative difference in odds of completing the study between participants with spouse and "other" as study partner is 2.37 (95%CI:1.18, 4.78, P-value:0.016).

## 1 Introduction

Research on prevention of Alzheimer disease usually requires long-term involvement of both participants and their partners. Advanced inference and prediction on the odds of subject completing the study increase the odds of a complete and valid analysis in the end.

In this study firstly I will draw inference on (i) the association between the partner type (child, spouse, friend and other) and the odds of completing the 48-month study and (ii) whether this association changes given different dementia status. Logistic regression model with some reasonably adjusted variables would be used to achieve this objective.

In order to analyze how adjusted variables relate to the response (whether completing the study or not), I categorize potential factors affecting the response into $2 \times 2$ folds: (subjective will vs. objective fact)$\times$(participant vs. partner). For instance, unwillingness to continue study is subjective while death and migration are objective. My following model building will be based on such perspective.

Secondly, I will build a predictive model to estimate the probability of completing the 48-month study. Model selection method is applied to acquire the best subset of variables that best fit the data. The criteria used for model selection is AIC, which measures the expected log-likelihood of the data given the model. To avoid over-fitting, I firstly consider linear form of variables. After determining the subset of variables, model diagnostics would examine such assumption. At last a goodness of fit test is performed to see whether the model is adequate. Model assessment through method of receiver operating characteristic (ROC) and area under the curve (AUC) will also be presented.

## 2 Methods

### 2.1 Source of the Data

The data was collected in a 4-year observational study from 39 sites scattered across the United States with 644 pairs of participants and partners. To some extent the data generalizes to the state though the sample size is not adequately large. However, because the data dose not contain any direct affecting information of why participants didn't complete the study, unmeasured confounding (e.g. death) can only be assessed by intervening factors (e.g. age).

At the baseline all participants were beyond 75 years of age with good health and no severe dementia. Demographical information provides potential subjective (education) and objective (age, gender) factors of both participants and partners. History of unhealthy habits and diseases may be used as potential objective factors. Cognitive ability evaluations including Clinical Dementia Rating Scale (CDRS), Modified Mini-Mental State Exam (mMMSE) and Cognitive Function Instrument (CFI) reflect the baseline dementia status of the participants, which may be good estimates for subjective factor, since participant with dementia might fail to remember the study.

No missing values among responses and variables of partner type. Major data missing is in variable of partner age (55 NA's) and much fewer missing in CFI (less than 7). I won't adjust partner age because it impairs the effect of partner type as a mediator. However, if partner age is a strong objective factor, then the missing is not ignorable, so further analysis needs to be addressed on assumption that partner age is not informative. As for missing in CFI, given the sample size, missing to this degree would have very little impact on analysis results, so further analyses presented were performed on a complete-case basis.

## 2.2 Statistical Methods

Logistic regression model with a logit link is applied due to the interest of odds and ability of interpretation. I use R for data inspection and model building. The model setting is as follows: Let $Y_i = 1$ and $Y_i = 0$ respectively denote the i-th participant completing the study or not, $\mu_i$ denote the expectation of $Y_i$ , $x_{ij}$ denote the value of j-th variable for i-th participant ($x_{i0} = 1$), and $\beta_j$ denote the regression coefficient for j-th variable. Then the Logistic regression model with a logit link is $log(\frac{\mu_i}{1-\mu_i}) = \sum_j^p x_{ij}\beta_j$. Then the interpretation of the exponentiated maximum likelihood estimate of $\beta_j, j > 0$, would be the relative difference in odds of completing the study between two groups of participants with one unit of difference in $x_j$ and similar in all other variables $x_k, k \neq j$.

For inference model, firstly I only adjust for parter type to see its marginalized effect. Then according to my previous perspective, I adjust for participant's age, gender and history of smoking and diseases as potential objective factors implying death (alcohol and drug abuse are not adjusted for their imbalanced distribution, see 3.1), and participant's education years as well as partner's as potential subjective factors implying willingness to complete the study. Notice participant's age may also implies partner type hence is expected to reduce confounding. I also adjust the categorized CDRS (= 0 vs. > 0) as a subjective factor. Other variables are not adjusted either for irrelevance to completing the study such as cognitive evaluation and partner's gender, or for impairing the effect of partner type such as partner's age, contact with participant and whether living together. Lastly, I add interaction between categorized CDRS and partner type to determine whether the association between response and partner type changes given different dementia status. Note that CDRS is chosen to describe dementia for it's explicitly defined compared to CFI and approximately consistent with mMMSE (shown in 3.1). In terms of partner type I choose "other" as the reference level so the exponentiated coefficient estimate with respect to the rest levels would be the relative difference in odds of completing the study between two groups of participants with partner type as corresponding level and "other" level, and similar in all other characteristics (age, gender, education etc.). 95% confidence interval and P value would be used to draw the final conclusion for the hypotheses. Since the response is binary, mean-variance assumption test is not necessary. Diagnosis of functional form and influential points will be performed. At last a Hosmer-Lemeshow goodness of fit test is presented using R function binary.gof() since sample size is not large and such method may avoid sparseness.

For predictive, I use a stepwise method initially with almost all variables in the model. Before conducting a stepwise procedure, I manually exclude partner age, for its vast missing values and little relevance to response (Figure 1). CFI variables are also excluded for their subjectivity and difficulty in modeling. History of alcohol and drug abuse are also not included for the same reason in the inference model. With the rest of variables all in the model, I use R function stepAIC(), which works by tentatively adding or excluding variables repeatedly to finally reach the model with lowest AIC, to acquire the best subset. Next I explore the interaction between variables in the selected subset, transformation of variables according to diagnostic analysis. A goodness of fit test is also presented for the predictive model. At last I use plot of ROC and AUC to assess the model prediction. Function loess() is used to fit sensitivity vs. 1-specificity data. Then method of Riemann sum is used to approximate AUC.

| Characteristic<br>Median (IQR) or N (%) | Total<br>N=644 | Complete<br>(418, 65%) | Incomplete<br>(226, 35%) |
|---|---|---|---|
| AGE | 79 (77, 82) | 78 (76, 81) | 79 (77, 83) |
| FEMALE | 375 (58%) | 246 (59%) | 129 (57%) |
| EDUCATION YEARS | 16 (12, 17.25) | 16 (13, 18) | 14 (12, 16) |
| SMOKING HISTORY | 250 (39%) | 159 (38%) | 91 (40%) |
| CVD HISTORY | 422 (66%) | 265 (63%) | 157 (69%) |
| CANCER HISTORY | 163 (25%) | 104 (25%) | 59 (26%) |
| mMMSE SCORE | 96 (93, 98) | 97 (94, 98) | 95 (91, 97) |
| CDRS = 0 | 401 (62%) | 279 (68%) | 122 (54%) |
| PARTNER TYPE | | | |
| SPOUSE | 270 (42%) | 180 (43%) | 90 (40%) |
| CHILD | 108 (17%) | 75 (18%) | 33 (15%) |
| FRIEND | 189 (29%) | 119 (28%) | 70 (31%) |
| OTHER | 77 (12%) | 44 (11%) | 33 (15%) |
| PARTNER AGE | 73 (58, 78) (55 NA) | 73 (58, 78) (41 NA) | 74 (58, 80) (14 NA) |
| MALE PARTNER | 164 (25%) | 109 (26%) | 55 (24%) |
| PARTNER EDUCATION | 16 (12, 17) | 16 (13, 17) | 14 (12, 16) |
| CONTACT FREQUENCY | 7 (4, 7) | 7 (4, 7) | 7 (4, 7) |
| COHABIT | 315 (49%) | 206 (49%) | 109 (48%) |

Table 1: Selected participant characteristics stratified by response.

# 3 Results

## 3.1 Descriptive Statistics

Table 1 presents some selected characteristics of interest for the 644 participants stratified by response. For continuous variables, medians and inter-quartile ranges are presented due to the heavy skewness of variables such as age, education years and mMMSE score. Here I abandon the history of alcohol and drug abuse due to their extremely imbalanced distribution (622 vs. 22 and 643 vs. 1 respectively). CFI is also omitted for issues addressed in 2.2. Note that the distribution difference is noticeable in historical and cognitive characteristics, as well as in different partner types. The issue of missing value is addressed in 2.1. Figure 1 shows that the distribution of partner age is not significantly different across response, implying probably little confounding in this variable. Finally figure 2 shows the approximately consistent trend between CDRS and mMMSE.

## 3.2 Model Summary

Below are logistic regression estimates from models addressed in 2.2 along with summary on 95% confidence interval (Table 2) and P-value (Table 3) from each one.

- Unadjusted model: The estimated relative difference in odds of completing the study between two groups of participants with partner type as "spouse", "child" or "friend" level, and "other" level, are 1.50 (95%CI:0.89, 2.52, P-value:0.124), 1.70 (95%CI:0.93, 3.14, P-value:0.086), 1.28 (95%CI:0.74, 2.19, P-value:0.377) respectively.

- Adjusted model: The estimated relative difference in odds of completing the study between two groups of participants with partner type as "spouse", "child" or "friend" level, and "other" level, and similar in age, gender, education of both participant and partner, history of smoking , medical history of cardiovascular disease and cancer, cognition status, are 1.38 (95%CI:0.79, 2.41, P-value:0.256), 1.66 (95%CI:0.88, 3.14, P-value:0.115), 1.26 (95%CI:0.72, 2.22, P-value:0.418) respectively.

- Model with interaction: The estimated relative difference of partner type effect to odds of completing the study between two groups of participants with CDRS=0 and CDRS>0, for "spouse", "child" and "friend" are respectively 3.91 (95%CI:1.29, 11.80, P-value:0.015), 3.20 (95%CI:0.88, 11.68, P-value:0.078), 2.83 (95%CI:0.89, 8.98, P-value:0.078), similar in age,

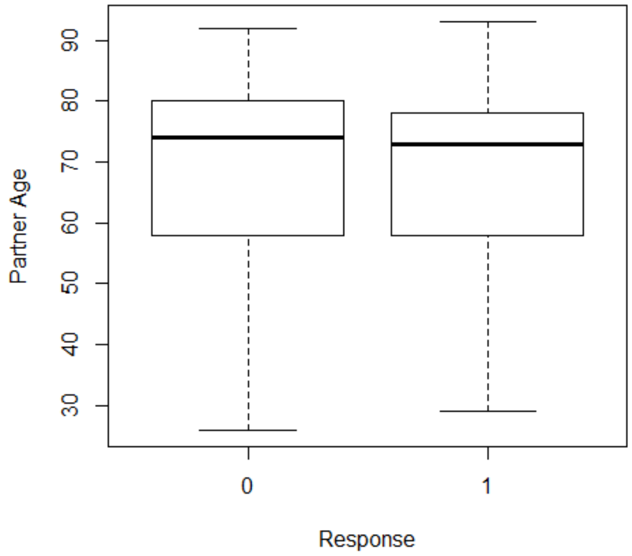

Figure 1: Partner age vs. Response.

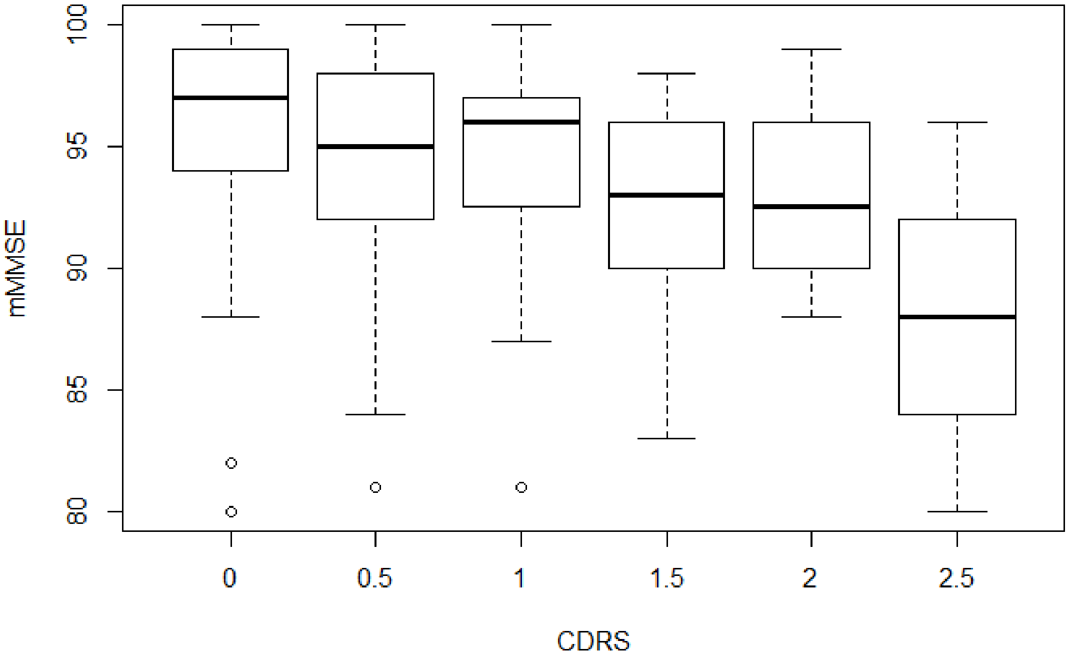

Figure 2: CDRS vs. mMMSE.

| Covariate<br>Est. (95% C.I.) | Unadjusted | Adjusted | Adjusted with Interaction |
|---|---|---|---|
| AGE | | 0.92 (0.88, 0.96) | 0.92 (0.87, 0.96) |
| FEMALE | | 1.16 (0.79, 1.70) | 1.18 (0.80, 1.73) |
| EDUCATION YEARS | | 1.09 (1.03, 1.16) | 1.09 (1.03, 1.16) |
| SMOKING HISTORY | | 0.88 (0.62, 1.26) | 0.87 (0.61, 1.25) |
| CVD HISTORY | | 0.86 (0.60, 1.23) | 0.86 (0.60, 1.24) |
| CANCER HISTORY | | 0.95 (0.64, 1.40) | 0.94 (0.63, 1.38) |
| CDRS=0 | | 1.61 (1.15, 2.27) | 0.56 (0.21, 1.47) |
| PARTNER TYPE | | | |
| SPOUSE | 1.50 (0.89, 2.52) | 1.38 (0.79, 2.41) | 0.61 (0.25, 1.46) |
| CHILD | 1.70 (0.93, 3.14) | 1.66 (0.88, 3.14) | 0.82 (0.30, 2.25) |
| FRIEND | 1.28 (0.74, 2.19) | 1.26 (0.72 ,2.22) | 0.67 (0.27, 1.66) |
| OTHER | Referent | Referent | Referent |
| PARTNER EDUCATION | | 1.03 (0.97, 1.10) | 1.03 (0.96, 1.09) |
| INTERACTION | | | |
| SPOUSE.CDRS=0 | | | 3.91 (1.29, 11.80) |
| CHILD.CDRS=0 | | | 3.20 (0.88, 11.68) |
| FRIEND.CDRS=0 | | | 2.83 (0.89, 8.98) |

Table 2: Selected logistic regression estimates for inference modeling

gender, education of both participant and partner, history of smoking , medical history of cardiovascular disease and cancer, cognition status.

Both unadjusted and adjusted models show no significant association between partner type and odds of completing the study. Age, participant's education and CDRS mostly explain the response. After adjusting for the interaction, only the effect of partner type "spouse" changes significantly with relative difference 3.91 (95%CI:1.29, 11.80, P-value:0.015). Since the association is not significant in CDRS=1 group, with relative difference in odds of completing the study as 0.61 (95%CI:0.25, 1.46, P-value:0.266), I still need to test the association in CDRS=0 group, which requires inference of linear combination of coefficients "SPOUSE" and "SPOUSE.CDRS=0". The result is that estimated relative difference in odds of completing the study between two groups of participants with both CDRS = 0, and partner type as "spouse" and "other", and similar in other adjusted variables, is 2.37 (95%CI:1.18, 4.78, P-value:0.016). At last a Hosmer-Lemeshow goodness of fit test with P-value 0.172 for adjusted model and 0.464 for model with interaction.

The selected variables through stepwise method based on AIC and their estimates are shown in Table 4. Further exploration of interaction terms indicates no more significant contribution to AIC, and diagnostics of linear form implying no need for variable transformation. A Hosmer-Lemeshow goodness of fit test with P-value 0.894 shows adequate fit to the data. Figure 3 shows the ROC curve for the predictive model and the calculated AUC through smoothing and Riemann sum approximation is 0.666.

## 4 Discussion

Result shows that given no strong evidence on dementia, participants with spouse as study parter is more likely to complete the study that those with person other than spouse, child and friend as study partner, under similar condition in other aspects. The analysis indicates weak association between the odds of completing the study and partner type, because if we slightly lower the confidence level, some of the results will become significant (e.g. P-value<0.1 child and friend as partner are more likely to complete the study given no strong evidence on dementia). Predictive model regards age, education, baseline dementia status and partner's memory as stronger predictors. So in general we could say younger participants with higher education and less sign of dementia, and with relative as study partner may be more likely to complete the study.

The limitation of the analysis is that we don't have accurate information to explain why participants stop coming, with which the conclusion of the analysis would be more certain. Hence further study should be more concerned about concrete reason why subjects fail to continue study. Another limitation of the analysis is insufficient use of the data with respect to intractable CFI

| Covariate<br>P-value | Unadjusted | Adjusted | Adjusted with Interaction |
|---|---|---|---|
| AGE | | 3.99e-4 | 2.31e-4 |
| FEMALE | | 0.437 | 0.400 |
| EDUCATION YEARS | | 5.14e-3 | 4.43e-3 |
| SMOKING HISTORY | | 0.489 | 0.457 |
| CVD HISTORY | | 0.407 | 0.417 |
| CANCER HISTORY | | 0.795 | 0.743 |
| CDRS=0 | | 6.16e-3 | 0.240 |
| PARTNER TYPE | | | |
| SPOUSE | 0.124 | 0.256 | 0.266 |
| CHILD | 0.086 | 0.115 | 0.704 |
| FRIEND | 0.377 | 0.418 | 0.384 |
| OTHER | Referent | Referent | Referent |
| PARTNER EDUCATION | | 0.336 | 0.409 |
| INTERACTION | | | |
| SPOUSE.CDRS=0 | | | 0.015 |
| CHILD.CDRS=0 | | | 0.078 |
| FRIEND.CDRS=0 | | | 0.078 |

Table 3: P-value of regression estimates for inference modeling

| Covariate | Est. (95% C.I.) P-value |
|---|---|
| AGE | 0.92 (0.88 ,0.97) 8.0e-4 |
| EDUCATION YEARS | 1.06 (1.00, 1.13) 0.052 |
| CDRS | 0.63 (0.44, 0.90) 0.012 |
| mMMSE | 1.08 (1.03, 1.14) 2.8e-3 |
| PARTNER REPORT DECLINE IN MEMORY | 0.67 (0.46 ,0.97) 0.032 |

Table 4: Selected logistic regression estimates for predictive modeling

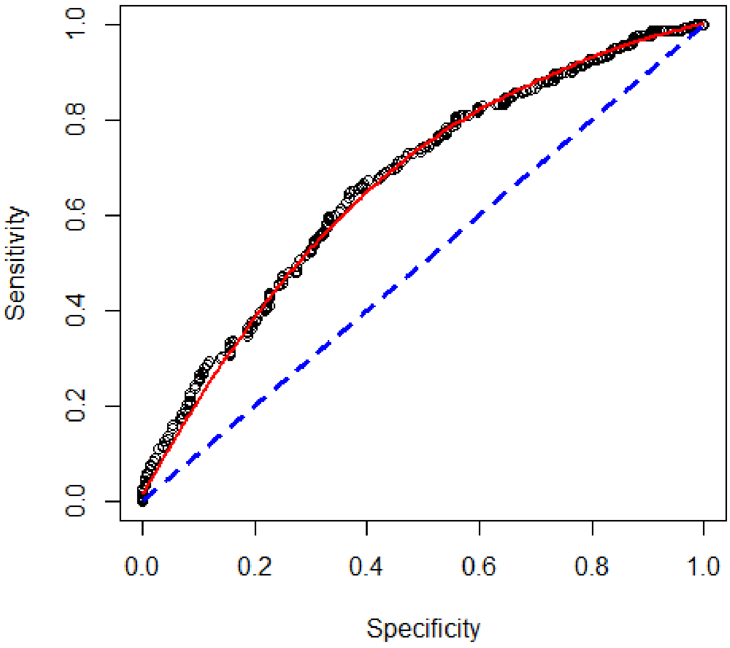

Figure 3: ROC curve for the predictive model

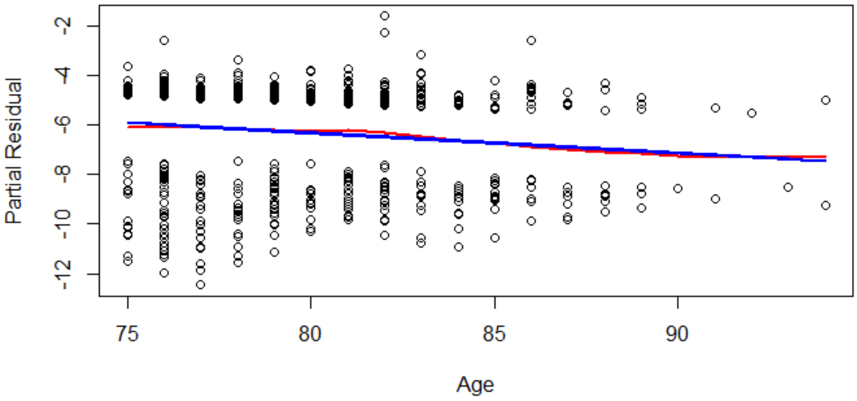

Figure 4: Parital residual vs. Age

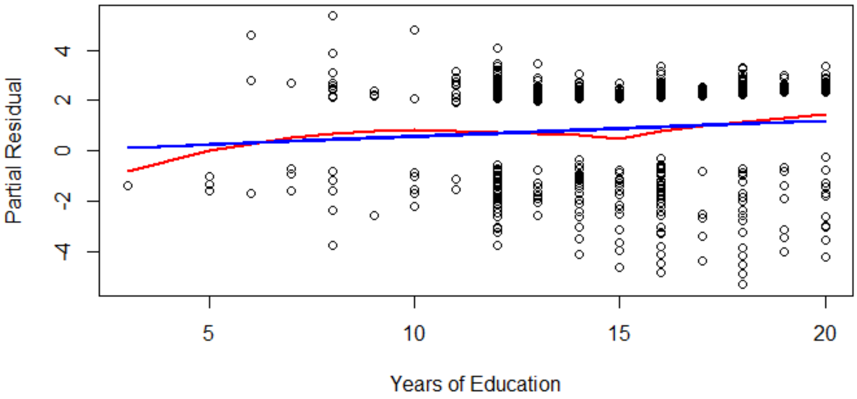

Figure 5: Parital residual vs. Years of Education

due to its subjectivity and multi-dimensionality. A systematic approach should be integrated by specialists on how to evaluate cognitive status through CFI.

# 5 Appendix

Model diagnostics include evaluating mean-variance assumption, linear form and influential points. Since the data response is binary and distribution could only be Bernoulli, it's not necessary to examine the mean-variance assumption. As for linear form assumption of variable, I use partial residual and data smoothing to see whether for each variable the linear relation would properly fit the data. For predictive model I examine three continuous variables of age, education and mMMSE in Figure 4, Figure 5 and Figure 6. As is shown that three smoothers fit the model pretty well so the linear assumption is appropriate. For inference model all continuous variables have been included above. At last I use Cook's distance to examine influential points. Figure 7 shows that no evident influential point in the dataset, similar pattern for both inferential and predictive models.

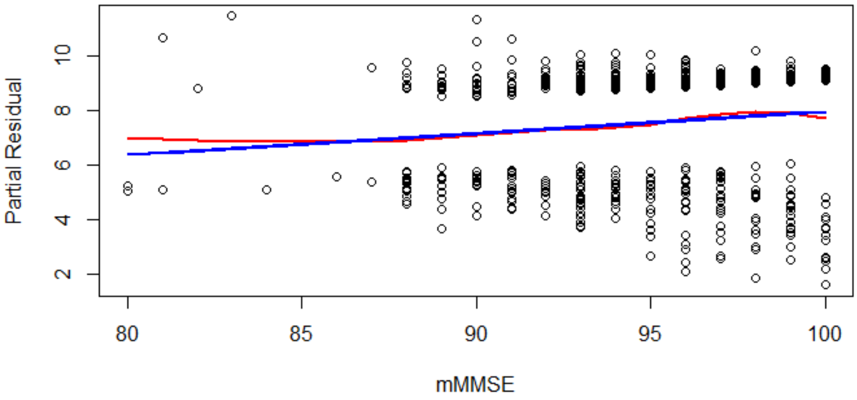

Figure 6: Parital residual vs. mMMSE

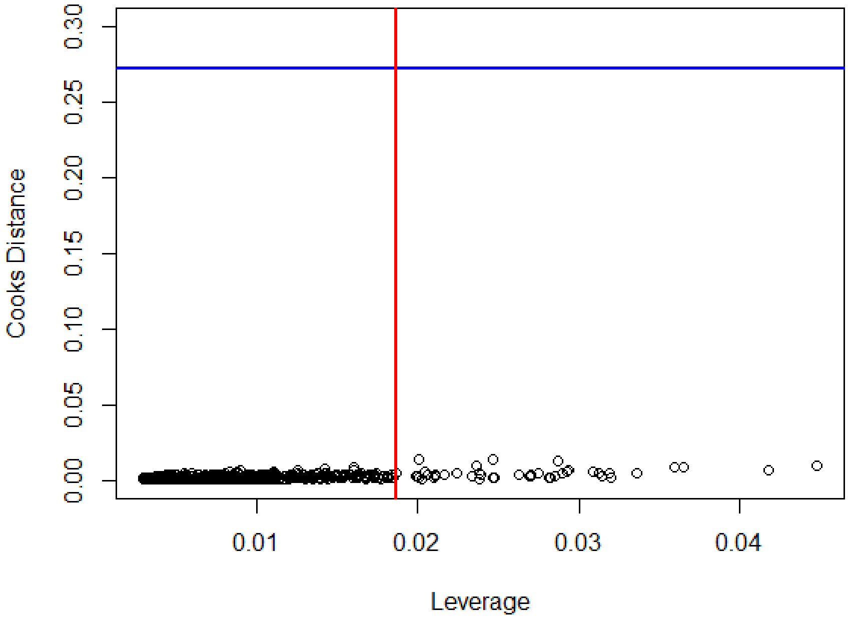

Figure 7: Inflential points diagnostics